AI 온라인 학습 도우미 PLUS

(Posco Learning Upgrade System)

포스코 AI 빅데이터 아카데미 9기 A1조
강민구 | 권수민 | 이용현 | 이정우 | 정해유

CONTENTS

1 프로젝트 소개
선정배경 및 유사기술

2 프로젝트 시연
시연 영상

3 개발 과정
활용 기술

4 기대효과
PLUS 의 활용방안

5 참고문헌
논문, 저널

1 프로젝트 소개 - 아이디어 선정배경

코로나 바이러스로 인해 온라인 강의 수요 증가

초등학교, 중학교, 대학교에 불문하고
인터넷을 이용한 원격 강의가 활발히 활용되고 있음

교육자와 학생 간의 쌍방향 소통이 어려운 문제

[NEWS1] 대학생 10명 8명 "온라인 강의 불편해 "

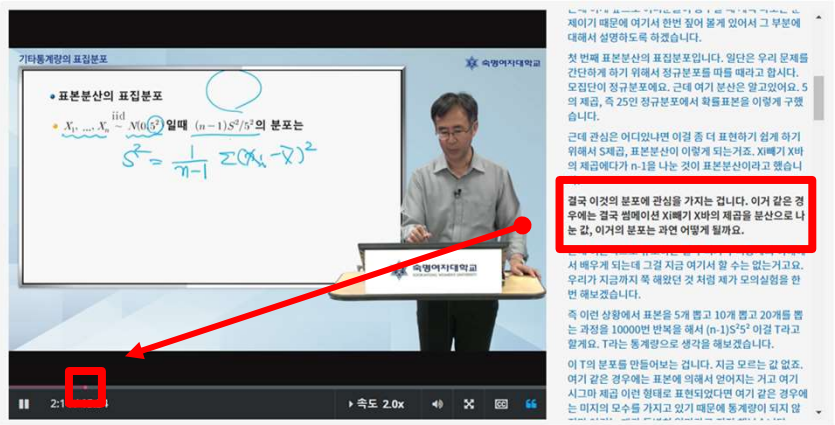
대학생들이 꼽은 온라인강의 단점으로는 '집중력 저하'(19.1%)가 가장 많았다. 이어서 △접속·서버 장애 불안정(16.6%) △온라인 강의 질 저하(16.2%) △수업 중 문답, 질의 처리가 어려움(13.4%) △팀 프로젝트 불가(6.8%) 등으로 답했다.

온라인 강의의 경우 질의 응답이 어렵다는 단점이 있음
이를 극복할 수 있는 플랫폼이 필요함.

" 교육자와 쌍방향 소통이 없더라도 학습에 불편함이 없도록 AI 학습 도우미를 만들면 어떨까? "

1 프로젝트 소개 - 유사기술

K-MOOC



영상 속 교사가 말하는 내용을 자막으로 제공

다시 듣고 싶은 내용을 클릭하면 해당 시간대로 이동

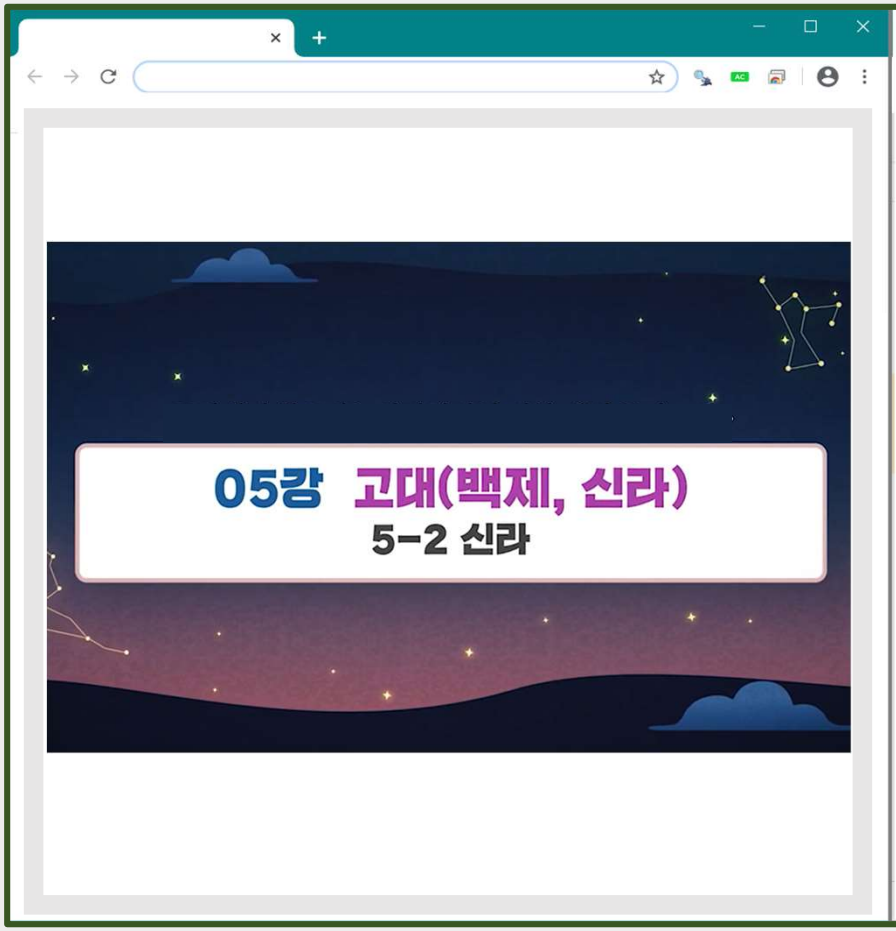
EBS-i

동영상 시청 중에 단어 검색이 가능하지만 사용자가

직접 검색 일반 검색 엔진을 사용하는 것과 큰 차별점은 없음.

단순 자막과 검색 기능은 제공하지만 사용자가 직접 내용을 찾고 검색해야한다.
이러한 문제를 극복하기 위해 **자동으로 중요한 키워드를 검색해주고,**
궁금한 내용을 자연어로 질문할 수 있는 플랫폼을 만들고자 한다.

1 프로젝트 소개 – 구체적인 사용모습



AI 학습 도우미 프로그램

실시간 자막 기능

실시간 교육 영상에서 교육자가 말하는 내용을 Text로 변환하여 보여주는 창이다. 교육자가 마지막으로 한 말은 굵은 표시를 한다.

자동 키워드 추출 & 검색 기능

실시간 자막 Text에서 생소하거나 어려운 단어를 추정하여 인터넷에서 검색한 결과 및 페이지 링크를 보여주는 창이다.

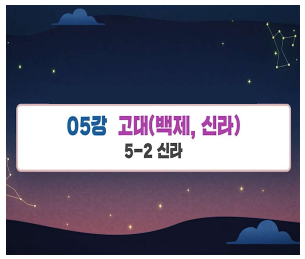
Q & A 기능

교육자 대신에 실시간 자막 Text를 기반으로 하여 교육생이 작성한 질문에 교육자 대신 AI의 대답을 보여주는 창이다.

2 프로젝트 시연

프로그램 실행

3 개발 과정 - 전체 구조도



온라인 강의

실시간 자막

질문 입력

Machine Reading Comprehension (QA)

QA MODEL

답변 추출

Named Entity Recognition (KEYWORDS)

NER MODEL

NER TAG 추출

키워드 추출

3 개발 과정 - Machine Reading Comprehension(QA)

- Machine Reading Comprehension(QA)이란?

  MRC(Machine Reading Comprehension)는 모델이 주어진
  지문(Context)를 학습하고 질문(Question)에 대한 답변을 추론하는 기술

[CONTEXT]

중세 고려를 이끌어 간 사람들은
<<호족>>이에요.바로 통일신라 말기에 등장했던
<<호족>>들이 중세 고려를 이끌어 가게 된 것이고,
대표적인 인물이 바로 태조 왕건이 되는 겁니다.

[QUESTION]

통일 신라 말기에 새롭게 등장한 세력은?

3 개발 과정 - Machine Reading Comprehension(QA)

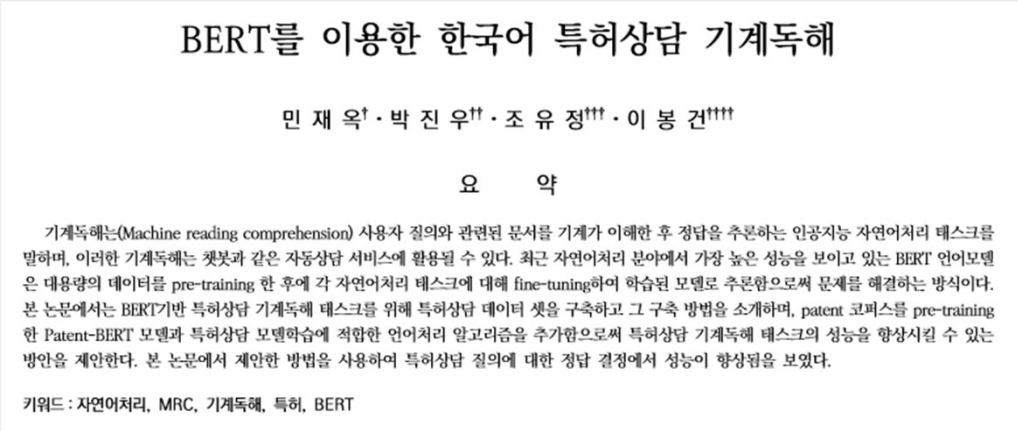
- 기계 독해: BERT 모델이 상위권 점유

Leaderboard

KorQuAD 1.0의 Test set으로 평가한 Exact Match(EM) 및 F1 score 입니다.

Rank	Reg. Date	Model	EM	F1
-	2018.10.17	**Human Performance**	80.17	91.20
1	2020.01.08	SkERT-Large (single model) Skelter Labs	87.66	**95.15**
2	2019.10.25	KorBERT-Large v1.0 ETRI ExoBrain Team	**87.76**	95.02
3	2020.01.07	SkERT-LARGE (single model) Skelter Labs	87.25	94.75
4	2019.06.26	LaRva-Kor-Large+ + CLaF (single) Clova AI LaRva Team	86.84	94.75
5	2020.01.03	SkERT Large (single model) Skelter Labs	87.28	94.66
6	2019.06.04	BERT-CLKT-MIDDLE (single model) Anonymous	86.71	94.55
7	2019.06.03	LaRva-Kor-Large + CLaF (single) Clova AI LaRva Team (LPT)	86.79	94.37

- QA 시스템에도 BERT 모델을 사용한 논문이 있음



BERT를 이용한 한국어 특허상담 기계독해

민 재 옥† · 박 진 우†† · 조 유 정††† · 이 봉 건††††

요 약

기계독해는(Machine reading comprehension) 사용자 질의와 관련된 문서를 기계가 이해한 후 정답을 추론하는 인공지능 자연어처리 태스크를 말하며, 이러한 기계독해는 챗봇과 같은 자동상담 서비스에 활용될 수 있다. 최근 자연어처리 분야에서 가장 높은 성능을 보이고 있는 BERT 언어모델은 대용량의 데이터를 pre-training 한 후에 각 자연어처리 태스크에 대해 fine-tuning하여 학습된 모델로 추론함으로써 문제를 해결하는 방식이다. 본 논문에서는 BERT기반 특허상담 기계독해 태스크를 위해 특허상담 데이터 셋을 구축하고 그 구축 방법을 소개하며, patent 코퍼스를 pre-training 한 Patent-BERT 모델과 특허상담 모델학습에 적합한 언어처리 알고리즘을 추가함으로써 특허상담 기계독해 태스크의 성능을 향상시킬 수 있는 방안을 제안한다. 본 논문에서 제안한 방법을 사용하여 특허상담 질의에 대한 정답 결정에서 성능이 향상됨을 보였다.

키워드 : 자연어처리, MRC, 기계독해, 특허, BERT

BERT(Bidirectional Encoder Representations from Transformers)
2018년 11월, Google에서 공개한 고성능의 언어모델

여러 가지 방법으로 QA 시스템을 구현할 수 있지만 기계독해에서 뛰어난 성능을 보여주며, 여러 논문을 참고할 수 있는 BERT 모델을 사용하기로 결정

3 개발 과정 - Machine Reading Comprehension(QA)

Pre-training

Original

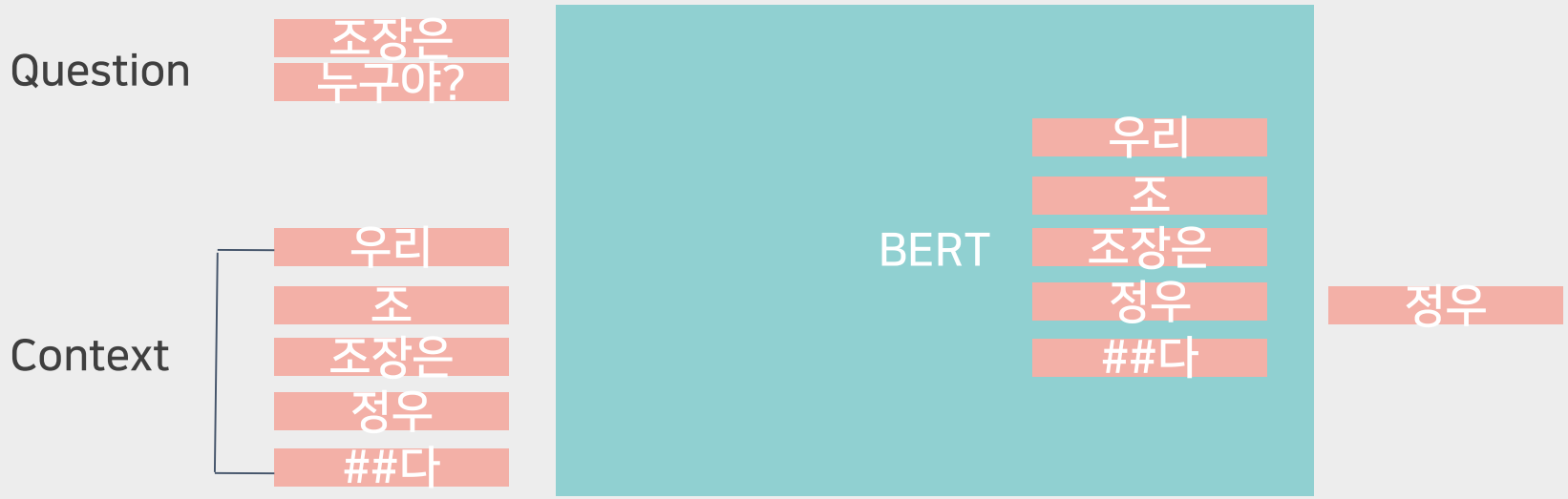
[CLS]	우리	조	조장은	정우	##다	[SEP]

Masking

[CLS]	우리	[MASK]	조장은	정우	##다	[SEP]

단어 중 일부를 [MASK] token으로 변환(15%)
[MASK] token을 predict 하는 pre-training 과정을 통해 BERT는 문맥을 파악하는 능력을 생성

Fine-tuning



Question에 정답이 되는 Paragraph의 substring을 뽑아내는 Fine-Tuning Task

3 개발 과정 - Machine Reading Comprehension(QA)

KorQuAD 1.0

- 한국어 MRC를 위해 만든 데이터셋
- 1,560개의 Wikipedia article에 대해 10,645 건의 문단과 66,181개의 질의응답 쌍으로 구성되어있음
- Training set: 60,407 개
- Test set: 5,774 개

[CONTEXT]

"2014년 12월 7일 토마스 바흐 IOC 위원장은 “8일부터 이틀간 열리는 IOC 총회에서 ‘어젠다 2020’이 확정되면 2018년과 2020년 동·하계 올림픽을 치르는 한국과 일본이 일부 종목을 분산 개최할 수 있다”고 말했다. IOC(국제올림픽위원회)는 2014년 12월 8일(한국 시간) 제 127회 총회에서 새로운 개혁안이 담긴 'Olympics Agenda 2020(어젠다 2020)'을 채택하였다. 새로운 개혁안을 통해 기존의 올림픽 개최 방식과 달리 국내 여러 도시들과 분산 개최가 가능하다.

[QUESTION]
"국제올림픽 위원회는 2014년 12월 8일 어젠다 2020을 채택하였는데 새로운 개혁안을 발표한 IOC 위원장의 이름은?"

[ANSWER] 토마스 바흐

3 개발 과정 - Machine Reading Comprehension(QA)

TEST DATA는 직접 제작

- KorQuAD TEST DATASET 으로 성능을 평가한 경우 F1 score는 93%
- 구어체로 구성된 자막 데이터를 기반으로 TEST DATASET 을 만들어서 성능을 평가함
- KorQuAD로 학습시킨 모델에 TEST DATASET을 입력한 후, 틀린 답을 내는 경우 해당 데이터를 기존 TRAIN DATASET에 추가하여 새로 모델을 학습시킴
- 이는 korquad 데이터가 문어체로 이루어져 있는데 본 프로젝트에서는 온라인 강의 텍스트인 구어체에 대한 학습을 시행하기 위함

[CONTEXT]

이 지도가 나오면 이건 나당전쟁을 의미하는 겁니다. 바로 나당전쟁의 승리를 거뒀던 장소입니다.매소성과 기벌포가 있어요. 그 위치를 잘 봐두세요. 매소성 전투와 기벌포 전투가 나오면이건 무조건, 무조건, 무조건 나당전쟁을 의미한다는 것을 기억하시면 되겠습니다. 됐죠? 신라는 결국 문무왕 때 삼국 통일을 완성하고 있더라. 삼국 통일에 앞장섰던 왕이 바로 문무왕이었습니다.결국 삼국 통일의 주인공은 신라. 누구도 예상하지 못했던 신라였다는 사실입니다.

[QUESTION] 나당전쟁의 승리를 거뒀던 장소는?
[ANSWER] 매소성과 기벌포

[QUESTION] 삼국 통일을 완성했던 왕은?
[ANSWER] 문무왕

3 개발 과정 - Machine Reading Comprehension(QA)

Q&A 결과

[CONTEXT]
어? 김부식 들어봤는데. 무슨 역사서를 쓴 사람이죠?
<<현존하는 우리나라에서 가장 오래된 역사서>>. 무엇?
<<삼국사기>>. 맞습니다. <<그 삼국사기를 썼던 김부식>>이
결국 이 묘청의 서경 천도 운동을 진압하게 되는데

[QUESTION]
현존하는 우리나라에서 가장 오래된 역사서 삼국사기를 기록한 인물은?
[ANSWER]
김부식

Test Accuracy	F1 Score
75.4% (전체 457문제 중 345문제 맞춤)	85.1%

*** F1 Score는 글자 단위(음절)로 해당 글자가 얼마나 겹쳐 나오는지 검토**

(예시) 대한민국 | 대한민국이다 → 8/10

3 개발 과정 - Machine Reading Comprehension(QA)

Q&A 결과

[CONTEXT]
<<호족들에 대한 전면적 숙청>>, 어마어마한 숙청을 했던 왕이 바로 이 <<광종>>입니다. 태조 왕건 때 보였던 왕권의 미약함. 이것을 강화시키기 위한 것이 광종의 목표였다는 것이죠. 그러기 위해서는 뭘 해야 돼요? 호족들을 숙청해야죠. 어떤 걸했냐면 먼저 <<노비안검법>>이라는 걸 시행합니다. 왜냐면 ...

[QUESTION]
광종이 호족 숙청을 위해 시행한 법은?
[ANSWER]
노비안검법

Test Accuracy	F1 Score
75.5% (전체 457문제 중 345문제 맞춤)	85.1%

* F1 Score는 글자 단위(음절)로 해당 글자가 얼마나 겹쳐 나오는지 검토

(예시) 대한민국 | 대한민국이다 → 8/10

3 개발 과정 - Named Entity Recognition (Keyword Extraction)

* 텍스트 내의 개체명의 의미를 파악하여 인명, 단체, 장소, 의학 등 어떤 유형에 속하는지 알아내는 모델
* 미리 정해진 tag 정보를 달아주는 정보 추출 task 중 하나. tag의 카테고리는 적용 domain에 따라 세분화된다.

ex) [그저께] [해유]는 [스타벅스]에서 커피를 한 잔 샀다.
=> [시간] [사람] [장소]

3 개발 과정 - Named Entity Recognition (Keyword Extraction)

개체명 예측을 위한 모델로 DistilBert + CRF를 선정

< Knowledge Distillation이란? >

- 복잡한 모델의 지식을 더 작고 가벼운 모델에 학습시키는 방법
- 12개의 Hidden layer를 이용한 무거운 Bert모델 대신,
  3개의 Layer를 가진 더 작은 Neural Network에
  지식을 학습시킨 모델인 DistilBert를 채택.
- 실시간으로 자막을 처리하여 신속하게 키워드를 추천해주기
  위해서는 가벼운 모델이 적절함.

	BERT (base)	DistilBERT
Parameter 수 (100만)	668	**410**
추론을 위한 시간 (초)	110	**66**
GLUE datasets 에 대한 Score	79.5	**77.0**

BERT와 DistilBERT의 성능 비교

< CRF (Conditional Random Field) 란? >

- CRF는 이전 태그 정보와 이후 태그 정보를 함께 이용하여 현재 태그를 예측함.
- 단일 값만을 이용하여 분류하는 softmax와는 다르게
  Sequence에 대한 정보를 반영하여 분류하게 되므로 성능 향상을 기대할 수 있음.

3 개발 과정 - Named Entity Recognition (Keyword Extraction)

DistilBert Model

KoBERT의 출력값을 soft target으로 두고,
같은 출력을 하도록 DistilKoBERT를 학습시킴

[_해 , 유 , 는 , _오 , 늘 , 도 , _포 , 스 , 텍 , 에, _갔 , 다]

KoBERT

Distil
KoBERT

Target tag = [PER_B, PER_I , PER_I , - , - , -
, ORG_B , ORG_I , ORG_I , ORG_I , - , -]

Output

Knowledge Distillation

CRF

DistilBert의 output을 CRF Layer에 입력하여
확률 벡터를 출력

[_해 , 유 , 는 , _오 , 늘 , 도 , _포 , 스 , 텍 , 에, _갔 , 다]

"포스텍"이 ' 단체'일 확률
⇒ 이전 단어인 "오늘도 " 와
이후 단어인 "갔다 " 를 고려하여 계산

[["_해"가 '사람'일 확률, "_해"가 '단체'일 확률 …] ,
["유"가 '사람'일 확률, "유"가 '단체' 일 확률 …] ,
…
["_다"가 '사람'일 확률, "_다"가 '단체' 일 확률 …]]

* KoBERT : BERT에 한국어 위키, 뉴스 등을 학습시켜 만든 한국어 버전 BERT. SKT에서 공개한 오픈소스.

3 개발 과정 - Named Entity Recognition (Keyword Extraction)

TF-IDF 알고리즘

- 정의

여러 문서로 이루어진 문서 군을 기준으로, 특정 문서에서 나타나는 단어가 문서 군 내에서 차지하는 중요도를 의미하는 통계적 수치로 문서 내에서 키워드 추출, 검색 결과 순위를 결정하는 용도로 사용됨

- 계산식

TF: 특정 단어가 문서 내에서 얼마나 자주 나타나는가

IDF: 문서빈도(DF)값의 역수

TF-IDF = TF x IDF

TF-IDF 와 NER 의 결합

- 예시

TF-IDF 를 통해 추출한 키워드
['노비안검법', '흑창' , '칭제건원' , '임시기구']

NER model 을 통해 추출한 키워드
(사람, 단체, 문명, 사건 개체명 추출)
['노비안검법', '권문세족', '태조']

두 키워드 집합의 교집합을 키워드로 추출

Keywords = ['노비안검법']

→ 키워드로 선정된 단어를 스크래핑을 통해 검색결과를 제공함

3 개발 과정 - Named Entity Recognition

네이버 NLP 데이터

- 네이버•창원대 NLP challenge 데이터 사용
- 총 1,063,571개

Index	Keyword	TAG
1	문재인	PER_B
2	일차	NUM_B
3	지방자치법	CVL_B
4	정보통신대학교	ORG_B
5	지난	DAT_B
6	라이벌전에	EVT_B
7	아메리카	LOC_B

웹 스크래핑으로 얻은 데이터

- 한국민족문화대백과사전에서 수집
- 총 7661개

키워드 한일합병

TAG EVT

문장 1910년 일제의 침략으로 한일합병조약에 따라 국권을 상실한 일.

키워드 분황사

TAG CVL

문장 분황사는 삼국시대 신라의 제 27대 선덕여왕 당시 창건한 사찰이다.

3 개발 과정 - Named Entity Recognition

1. 한국민족문화대백과사전에서 한국사 관련 문장 스크래핑

1910년 일제의 침략으로 한일합병조약에 따라 국권을 상실한 일

2. 네이버 데이터로 학습시킨 Model (acc : 95%) 에 문장 입력

1910년 일제의 침략으로 한일합병조약에 따라 국권을 상실한 일

날짜 지명 < 지명 > 사람

3. 정답 label 설정

1910년 일제의 침략으로 한일합병조약에 따라 국권을 상실한 일

날짜 지명 < 단체/회담 > 사람

일반적인 데이터에 대한 성능이 95%이므로 한국사에 대한 단어만 잘 맞추지 못한다고 가정

=> 다른 단어가 틀렸더라도 target 단어(한일합병조약)에 대한 태그만 수정하여 이를 정답 label로 가정

3 개발 과정 - Named Entity Recognition

키워드 추출 결과

[CONTEXT]
대농장을 몰수해서 나눠 주려는 그런 모습들도 보이고 있죠. 이 모습들이,
기억나야 될 게 광종 기억나요? 공민왕은 <<전민변정도감>>을 실시했습니다.
제가 왕 중에서 제일 시험에 많이 나오는 게 고려 전기에는 광종이 있다고
그랬죠? 고려 후기에는 공민왕이에요.

[KEYWORDS]
전민변정도감
[ANSWER]
고려 후기 권세가에게 점탈된 토지, 농민을 되찾기 위해 설치된 임시관서

Train Accuracy	Test Accuracy
96.1%	90.1%

Category	PER_B	PER_I	ORG_B	ORG_I	CVL_B	CVL_I	EVT_B	EVT_I
F1 score	0.89	0.82	0.87	0.74	0.85	0.48	0.81	0.7

3 개발 과정 - Named Entity Recognition

키워드 추출 결과

[CONTEXT]
그런데 금나라가 점점 커지는 거야. 엄청나게 커집니다. 이렇게 되면서 더이상은 그들과 맞서기에는 부담스러운 위치까지 간 거죠. 이때 사대를 요구하자 당시 실권자였던 이자겸과 김부식. 어? 느낌 오죠, 여러분들. 배웠잖아, 우리. 앞에서 <<문벌귀족>> 시대 모순의 중심에 있었던 인물들.

[KEYWORDS]
문벌 귀족
[ANSWER]
대대로 내려오는 그 집안의 사회적 신분이나 지위

Train Accuracy	Test Accuracy
96.1%	90.1%

Category	PER_B	PER_I	ORG_B	ORG_I	CVL_B	CVL_I	EVT_B	EVT_I
F1 score	0.89	0.82	0.87	0.74	0.85	0.48	0.81	0.7

4 기대 효과

학습자의 학습능률 향상

1. 학습자가 질문하면 수업내용을 기반으로 QA 기능을 제공

2. 수업과 관련된 핵심 키워드 추출, 백과사전 검색 결과 제공

3. 자막을 통해 놓친 수업내용을 다시 볼 수 있음

플랫폼 걱정 NO! 어떤 강의든 OK!

다양한 플랫폼 (ZOOM, 행아웃)을 사용하는 원격 수업에서 활용 가능

교육자의 수업 흐름 유지

수업의 흐름을 끊을 수 있는 기초적인 질문들을 키워드 추출 기능과 질문 기능을 통해 해결 가능

5 참고문헌

기계 독해 관련 문헌

Jacob Devlin외 3명, 2018, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", "Cornell University"
민재옥 외 3명, 2020, "Korean Machine Reading Comprehension for Patent Consultation Using BERT", 정보처리학회논문지

개체명 인식 관련 문헌

VictorSANH 외 3명 , 2020 , "DistilBERT,a distilled version of BERT : smaller, faster,cheaperandlighter" , NeurIPS'19
이한동, 2017, "단어 간 연관성을 고려한 키워드 추출 알고리즘", 숭실대 소프트웨어특성화 대학원

Thank You

포스코 AI 빅데이터 아카데미 9기 A1조
강민구 | 권수민 | 이용현 | 이정우 | 정해유

0 첨부자료

개체명 분류 카테고리

	개체명 범주	태그	정의
1	PERSON	PER	실존, 가상 등 인물명에 해당하는 것
2	FIELD	FLD	학문 분야 및 이론,법칙,기술 등
3	ARTIFACT WORDS	AFW	인공물로 사람에 의해 창조된 대상물
4	ORGANIZATION	ORG	기관 및 단체와 회의/회담을 모두 포함
5	LOCATION	LOC	지역명칭과 행정구역 명칭 등
6	CIVILIZATION	CVL	문명 및 문화에 관련된 용어
7	DATE	DAT	날짜
8	TIME	TIM	시간
9	NUMBER	NUM	숫자
10	EVENT	EVT	특정 사건 및 사고 명칭과 행사 등
11	ANIMAL	ANM	동물
12	PLANT	PLT	식물
13	MATERIAL	MAT	금속, 암석, 화학물질 등
14	TERM	TRM	의학 용어, IT 관련 용어 등 일반 용어를 총칭

조원소개

TITLE: 전국에 흩어진 a1조

(정우) 기계독해 QA, 데이터 전처리 및 BERT 연구

(용현) 기계독해 QA, 데이터 전처리 및 BERT 연구

(수민) NER 및 TF-IDF를 통한 키워드 추천

(해유) NER을 위한 DistilBert 및 CRF에 대한 연구

(민구) GUI 개발 및 시스템 통합